

جامعة دمشق

كلية الهندسة المعلوماتية



**Airline Passenger Satisfaction**

2 / 7 / 2021

مهند الطباع محمد علاء خير الله

محمد سامي العش

**المخلص**

**الملخص:**

تحتوي مجموعة البيانات على استبيانات حول رضا المسافرين في شركات الطيران. والهدف منها هو تصنيف الزبائن إلى قسمين:

1. زبائن راضية عن الشركة.
2. زبائن غير راضية أو طبيعية.

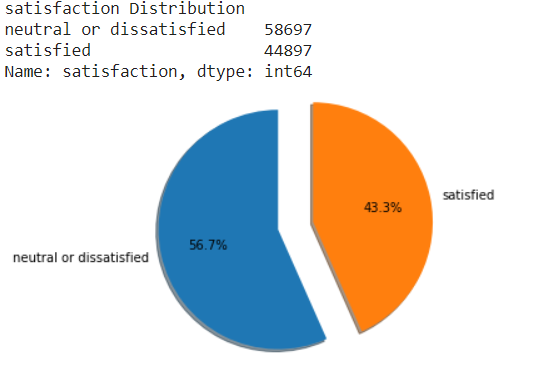
**مجموعة البيانات:**

تحتوي البيانات على 24 عمود (features) وعمود واحد كهدف (target) وعدد الأمثلة 103903.

**توصيف الأعمدة:**

1. رقم الراكب ضمن البيانات.
2. رقم التعريف الخاص بالراكب (ID).
3. الجنس: جنس الركاب (أنثى، ذكر).
4. نوع العميل: نوع العميل (عميل مخلص، عميل غير مخلص).
5. العمر: العمر الراكب.
6. نوع السفر: الغرض من رحلة الركاب (سفر شخصي، سفر عمل)
7. الدرجة: درجة السفر في طائرة الركاب (Business، Eco، Eco Plus).
8. مسافة الرحلة: مسافة الرحلة لهذه الرحلة.
9. خدمة wifi على متن الطائرة: مستوى الرضا عن خدمة wifi على متن الطائرة (0 -> 5).
10. وقت المغادرة / الوصول مناسب: مستوى الرضا عن وقت المغادرة / الوصول مناسب.
11. سهولة الحجز عبر الإنترنت: مستوى الرضاء عن الحجز عبر الإنترنت.
12. موقع البوابة: مستوى الرضاء لموقع البوابة.
13. الطعام والشراب: مستوى الرضاء عن الطعام والشراب.
14. الصعود عبر الإنترنت: مستوى الرضا عن الصعود عبر الإنترنت.
15. راحة المقعد: مستوى الرضا عن راحة المقعد.
16. الترفيه على متن الطائرة: مستوى رضا عن وسائل الترفيه على متن الطائرة.
17. الخدمة على متن الطائرة: مستوى الرضا عن الخدمة على متن الطائرة.
18. خدمة غرفة الساقين: مستوى الرضا عن خدمة غرفة الساقين.
19. تعامل مع الأمتعة: مستوى الرضا عن التعامل مع الأمتعة.
20. خدمة تسجيل الوصول: مستوى الرضا عن خدمة تسجيل الوصول.
21. الخدمة على متن الطائرة: مستوى الرضا عن الخدمة على متن الطائرة.
22. النظافة: مستوى الرضا عن النظافة.
23. تأخير المغادرة في دقائق: تأخير الدقائق عند المغادرة.
24. تأخير الوصول في دقائق: تأخير الدقائق عند الوصول.
25. الرضا: مستوى الرضا عن شركات الطيران (الرضا أو الحياد أو عدم الرضا).

**تنظيف البيانات**



إن كمية البيانات التي لدينا غير متوازنة ولكن وجود كمية بيانات كبيرة

يسمح بالتخلص من جزء من البيانات الموجودة في الصنف الأكبر من البيانات.

قبل القيام بالتخلص من البيانات المسببة للانحياز سيتم خلط البيانات كي يتم

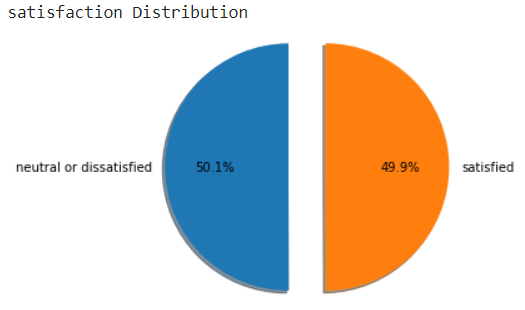
التأكد من أن البيانات التي تم الاحتفاظ بها تشمل كل أجزاء البيانات لأنه من

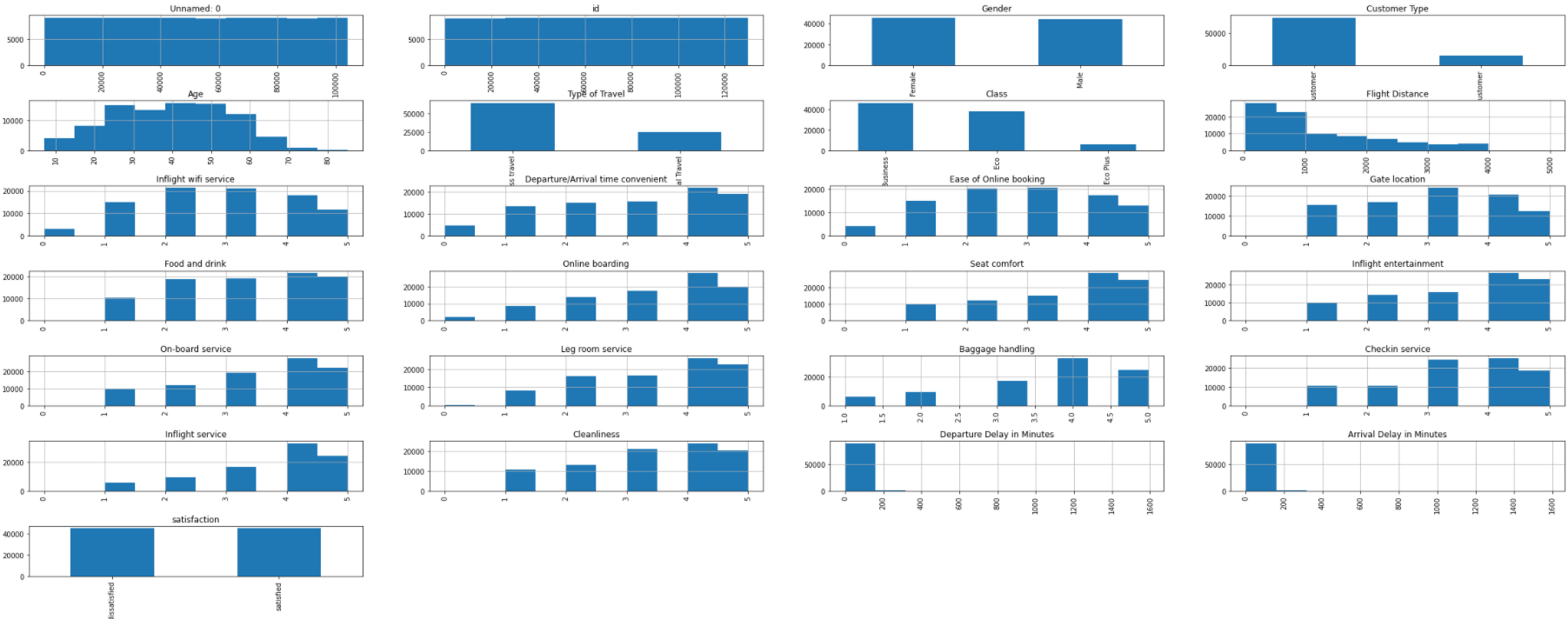
الممكن أن تكون البيانات موزعة بطريقة ذات معنى وعند القيام بالتخلص من قسم منها دون الخلط سيكون قد تم خسارة بيانات مهمة.

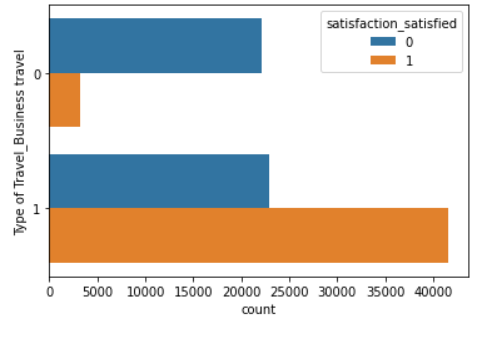
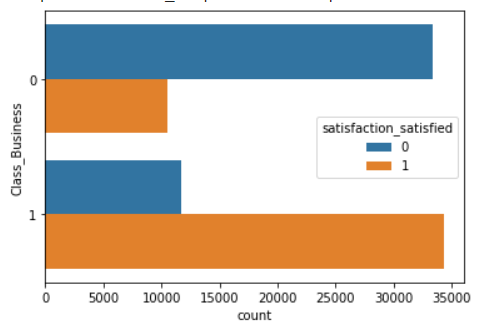
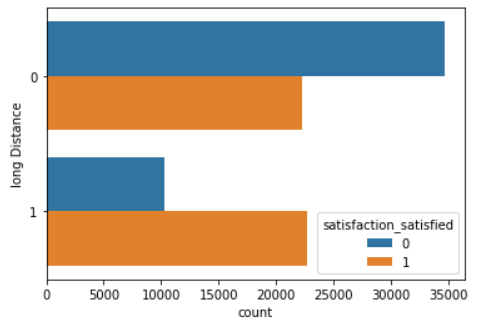
كما أنه يجب التخلص من الأسطر التي تحوي قيم غير ممتلئة (nan).

ملاحظة: يوجد حل أخر بدل القيام بالتخلص من البيانات الفارغة يتم ملأها بقيم تمثل المتوسط أو القيم الوسطى ولكن لا حاجة لذلك في هذه المسألة لأن كمية البيانات المتوفرة كبيرة وكافية.

تم التخلص من الأعمدة التي ليس لها معنى في هذه المسألة (رقم الراكب ضمن البيانات – الرقم التعريفي الخاص بالراكب).

وتصبح للبيانات تحتوي على 22 عمود (features) وعمود واحد كهدف (target) وعدد الأمثلة 89897. 

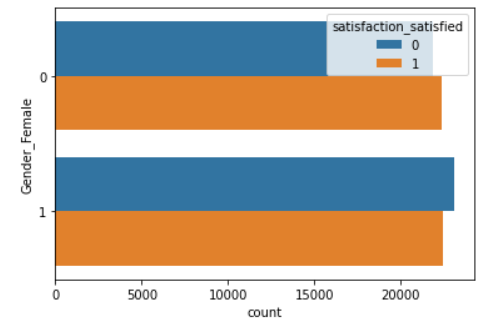
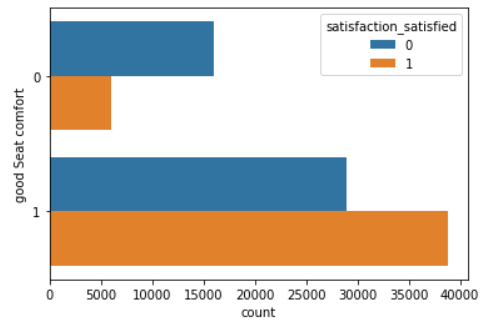
**تحليل البيانات**

****

أغلب الزبائن الذين سافروا مع الشركة لمسافات طويلة يشعرون بالرضا وذلك على عكس الزبائن الذين سافروا مع الشركة لمسافات قصيرة.

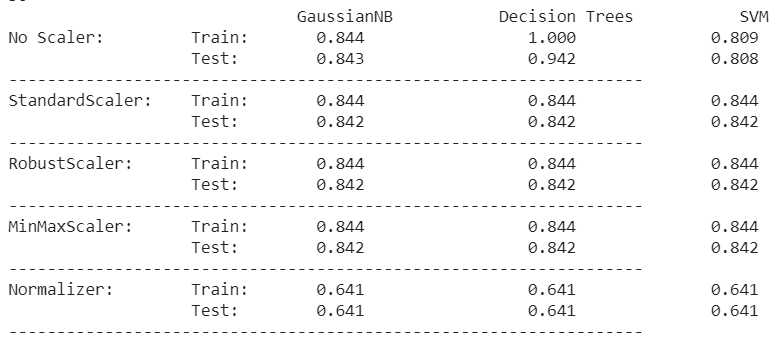
يجب على الشركة إعطاء اهتمام أكبر واجراء تحسينات على الدرجات التي ليست درجة أعمال لأن معظم الزبائن الذين ليسوا على درجة الأعمال لا يشعرون بالرضا

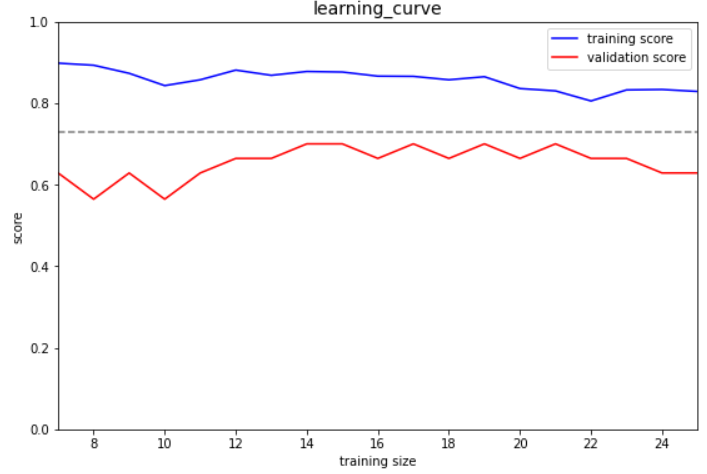
يوجد علاقة قوية بين سبب السفر ورضا الزبون فالزبائن الذين سافروا للعمل كانوا أكثر رضا من الزبائن الذين سافروا لأسباب شخصية.

****

المقولة التي تقول إن النساء أكثر تذمراً من الرجال غير صحيحة فلا وجود لعلاقة بين جنس الراكب ورضاه.

يجب على الشركة تحسين مقاعد الطائرة لأن الزبائن الذين لم يكونوا يشعروا بالراحة بسبب المقاعد يشعرون بعدم الرضا هن الشركة





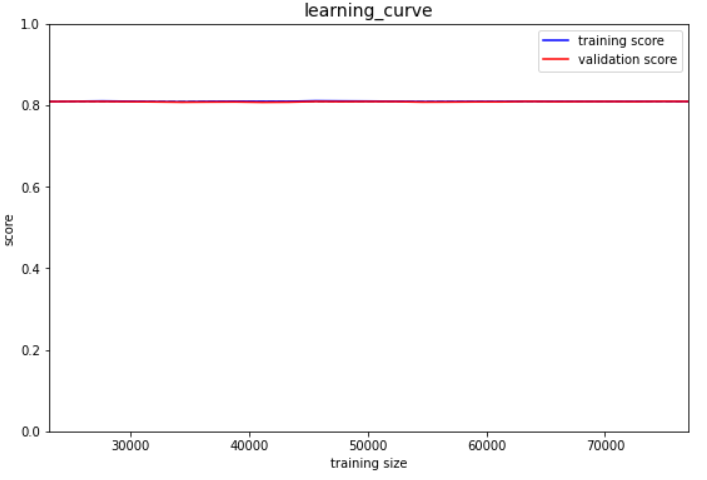
رسم توضيحي (1)

الشكل مكبر من أجل أول 25 مثال

من الشكل البياني يظهر لدينا وجود فائَض في البيانات ويمكن التخلص من قسم منها ولكن التخلص من قسم لن يحسن الأداء ولكنه فقط يسبب زمناٌ أكبر بالتدريب لذلك لا بأس بالاحتفاظ بها وعدم التخلص من أي جزء منها.

**منحنيات التعلم**

يجب التأكد من أن كمية البيانات التي تم التدريب عليها كانت كافية.



**التحويل إلى الشكل الرقمي**

لا يمكن العمل على البيانات بشكلها الحالي لأنه هنالك عدد من الأعمدة تحتوي على سلاسل محرفيه وبجب تحويلها إلى الشكل الرقمي.

يوجد عدة طرق للتحويل لكن أهمها.

1. إعطاء رقم لكل خيار ممكن ضمن هذا العمود (أي تحويلها إلى Label) وهنا تظهر لدينا مشكلة أن البيانات الرقمية لها دلالات ويتم التعلم خلال عملية التدريب هذه الدلالات ومثال على ذلك عند القيام بتحويل عمود الجنس باستعمال هذه الطريقة وفي حال تم إعطاء الذكر الرقم واحد والأنثى رقم اثنان سيتم تعلم العلاقة أن كل أنثى تساوي ذكرين وهذا الأمر خاطئ.
2. تحويل كل عمود إلى أعمدة بعدد الصفات الفريدة ضمن هذا العمود وتدعى هذه الطريقة ب One-Hot-Encoding (Dummy variables) وتتميز هذه الطريقة أن عملية الترميز تمنع استنتاج علاقات بين العناصر.

تم استعمال الطريقة الثانية لتصبح للبيانات تحتوي على 27 عمود (features) وعمودان كهدف (target) وعدد الأمثلة 89897.

إن العمودان الهدف يوجد بينهما ارتباط تام عكسي لذلك يمكن الاستغناء عن أحدهما.

### اختيار النماذج

هنالك أنواع من الخوارزميات تتأثر بتوزع البيانات لذلك سنقوم بإجراء مجموعة من التحويلات على البيانات قبل القيام بالتدريب،

هذه التحويلات هي كالتالي:

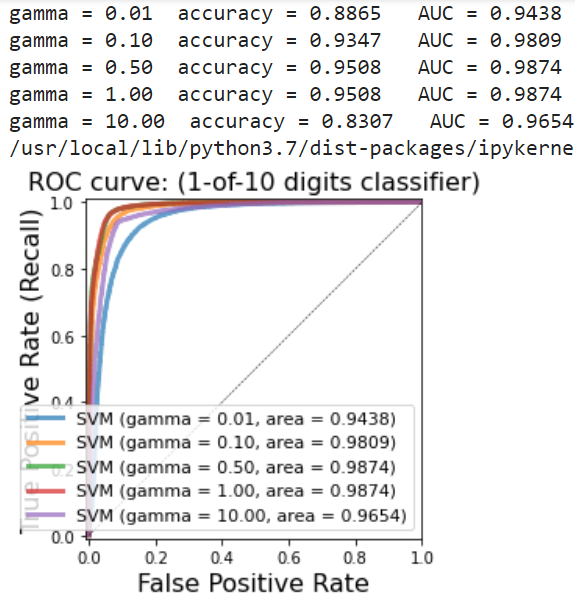
StandardScaler RobustScaler,  MinMaxScaler,  Normalizer,.

سنقوم بالتدريب على ثلاثة أنواع من الخوارزميات: GaussianNB, DecisionTreeClassifier, SVC.

ظهرت ليدنا النتائج الظاهرة في الرسم التوضيحي (1) وذلك بعد القيام بعملية تقسيم للبيانات إلى قسم للتدريب وقسم للاختبار.

هذا الشكل هو تكبير للشكل الأول لأول 25 مثال

|  |  |
| --- | --- |
| 93.88 | Decision Tree |



**roc curve**

كلما كان الرسم أقرب للواحد كلما كان أفضل.

### اختيار الميزات

تم اختيار ثلاث طرق:

1. SelectPercentile وتم تخفيض الأبعاد إلى 13 فكانت الميزات التي تم الحفاظ عليها هي كما يلي:

Flight Distance, Inflight wifi service, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Cleanliness, Type of Travel\_Business travel, Type of Travel\_Personal Travel, Class\_Business, Class\_Eco

وهي تطابق نتائج التحليل الموضح في فقرة تحليل البيانات

النتائج كما يلي:

|  |  |
| --- | --- |
| 70.05 | SVM |
| 83.92 | GaussianNB |
| 94.17 | Decision Tree |

1. RandomForestClassifier وتم تخفيض الأبعاد إلى 14 فكانت النتائج كما يلي:
2. RFE وتم تخفيض الأبعاد إلى 10 فكانت النتائج كما يلي:

Decision Tree:

SVM:

تم دراسة العلاقات بين كل الأعمدة النظر في مدى ترابط هذه الأعمدة.

**المراجع:**

* Support Vector Machines: Theory and Applications
* Short Survey on Naive Bayes Algorithm
* Random Forests and Decision Trees.

